MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media

CHARACTER SETS AND ENCODING OPTIONS

December 2007

CONTENTS

DISCLAIMER

INTRODUCTION

MARC 21 records intended for broad, standard interchange must employ either of two character encoding schemes. Only one of them may be used within a single record. The encoding now known as MARC-8 was introduced in 1968 with the beginning of the use of the MARC format. Over the years it has grown to include code points for a large repertoire of characters including Latin, Cyrillic, Arabic, Hebrew, and Greek scripts and over 15,000 characters used in writing Chinese, Japanese and Korean. The MARC-8 encoding is derived primarily from a collection of international standard character sets. These are identified in Part 2. The total collection of characters that can be represented in MARC-8 encoding is called the MARC-8 character repertoire. This extensive repertoire is adequate for many library environments. No further additions will be made to it.

Alternatively, the Universal Character Set (UCS or ISO/IEC 10646) encoding may be used. Its first version was published in 1993. As the name implies, the UCS aims to provide, in a single system, code points for the characters of all written languages. At present it includes over 100,000 characters used in dozens of scripts. ISO/IEC 10646 was developed in conjunction with the Unicode Consortium, an international group of industries, educational institutions, government agencies, etc. The consortium provides the primary energy for maintenance and expansion of the UCS. For that reason the UCS is frequently called Unicode. In this specification the terms UCS/Unicode, UCS, and Unicode may be considered synonymous when referring to the standard, either as encoding or as repertoire.

With the constantly growing adoption of the UCS/Unicode standard it will become a preferred option also for libraries. Conversions to Unicode have already taken place in many large library systems. When UCS/Unicode encoding is used in MARC 21, characters are expressed in the UCS transformation format, UTF-8. More information is given in Part 3.

- Part 1 provides guidelines for character set handling in MARC 21 records that is common to both the MARC-8 and UCS/Unicode encoding environments.
- Part 2 specifies the handling of character sets within the MARC-8 environment.
- Part 3 describes encoding in the UCS/Unicode environment.
- Part 4 specifies the issues involved in converting back and forth between the MARC-8 environment and repertoire and the UCS/Unicode environment and repertoire.
- Part 5 specifies, in the form of code tables, the MARC-8 repertoire and its encodings.

DEFINITIONS

Italicized terms found within definitions are terms for which definitions are also provided.

ASCII.
Acronym for American Standard Code for Information Interchange (ANSI X3.4), a 7-bit coded character set used as the default in MARC-8 encoding and, in its international counterpart (ISO/IEC 646 (IRV), serving as the foundation of the Universal Character Set (UCS). Consequently, *code points* less than 80 (hex) have the same meaning in both of the encodings used in MARC 21 and may be referred to as ASCII in either environment. It is useful to identify various subsets of the ASCII repertoire that are referenced in MARC 21 documentation:

ASCII numerics.
ASCII code points 30(hex) through 39(hex).

ASCII uppercase alphabetics.
ASCII code points 41(hex) through 4F(hex) and 50(hex) through 5A(hex).

ASCII lowercase alphabetics.
ASCII code points 61(hex) through 6F(hex) and 70(hex) through 7A(hex).

ASCII graphic symbols.
The ASCII graphic characters other than numerics, alphabetics, space, and delete. Code points 21(hex) through 2F(hex), 3A(hex) through 3F(hex), 40(hex), 5B(hex) through 5F(hex), 60(hex), and 7B(hex) through 7E(hex) are included.

ASCII graphics.
All ASCII characters (including space, numerics, alphabetics and graphic symbols) found in positions 20(hex) through 7E(hex).

ASCII space.
ASCII point 20(hex), an atypical graphic characterized by the lack of a written symbol. It has the unique property of being recognized by the standard non-ASCII graphic character sets employed in MARC-8 even though 20(hex) is not defined in those sets.

ASCII delete.
ASCII code point 7F(hex), a control character never used in MARC 21.

base character.
A graphic character that is not a *combining character*, but one with which one or more combining characters may be associated.

bidirectional script.
A *script* in which the primary display direction is conventionally reversed in specific situations. The most commonly encountered examples are the Arabic and Hebrew scripts, written from right-to-left in general but displaying multi-digit numbers left-to-right.

bit.
Short for binary digit. One of the two digits in a base 2 number system. Conventionally these are represented by 0 and 1.

byte.
A sequence of consecutive *bits* addressed and interpreted as a group. In current usage it is understood to contain eight bits unless otherwise qualified. An 8-bit byte is also called an *octet*.

character.
A unit of information used for the organization, control or representation of textual data.

coded character set.
A collection of characters in which each has been assigned a numeric *code point*. In this document, a reference to a character set assumes a coded set.

code extension.
The techniques for encoding *characters* that are not included in a given *coded character set*.

code point.
Any integer in a particular *codespace*.

code table.
A list or matrix identifying the *character* allocated to each *code point* in a *coded character set*.

codespace.

A range of integers available for encoding characters. The Unicode codespace includes integers from 0 to 10FFFF(hex). The codespaces of MARC-8 character sets, other than the East Asian Character Code, are limited to integers between 0 and FF(hex).

combining character; combining mark.

A character representing a mark, point, or sign used in conjunction with alphabetic or other graphic characters to distinguish them in form, sound, or meaning (usually intended to be displayed above or below an alphabetic graphic character).

control character.

A *control function* that is coded as a single *code point*.

control function.

An action that affects the recording, processing, transmission or interpretation of data and that has a coded representation consisting of one or more *code points*.

diacritical marks; diacritics.

A subset of the *combining characters*, but in common usage synonymous with the broader term.

escape (ESC).

A *control character* (ASCII 1B(hex)) which is used to provide additional *characters* by *code extension*. It alters the meaning of a limited number of contiguously following encoded characters, which form an *escape sequence*.

escape sequence.

A byte string that is used to *invoke* a new *working set* in *code extension* procedures. It comprises two or more characters, of which the first is the *escape* character.

field orientation.

Refers to the direction that *graphic characters* in a field are intended to be displayed and read (e.g., from left to right, or from right to left). In a MARC 21 record, the characters are to be recorded in their logical order, from the first character to the last character, irrespective of the direction they are intended to be read.

field orientation code.

A code that indicates the direction in which the displayed or printed *graphic characters* of a field would have been written and are intended to be displayed and read.

final character.

The *character* that terminates an *escape sequence*.

graphic character.

A *character*, other than a *control character*, that has a visual representation normally handwritten, printed, or displayed.

hexadecimal; hex.

Referring to a number system with sixteen digits, usually represented by 0-9 and A-F, each of which corresponds to a pattern of four bits. Hexadecimal notation is widely used for expressing the *scalar values* of *code points* and other numeric values. It is especially useful where *octets* are important because an octet can be expressed as two hex digits.

intermediate character.

Any *character* in an *escape sequence* occurring between the *escape* character and the *final character*.

invoke.

To designate a *coded character set* as the set of *code points* to be used in interpreting data.

MARC-8 encoding.

In this document MARC-8 encoding refers to character set encodings of the MARC-8 repertoire as described in Part 2 and specified in Part 5.

MARC-8 repertoire.

Over 16,000 characters for Latin, Cyrillic, Arabic, Hebrew, and Greek scripts and Chinese, Japanese and Korean ideographs, etc. as described in Part 2 and defined in Part 5 of this document.

nonspacing graphic character.

In this specification, the term is synonymous with *combining character*.

octet.

A group of eight consecutive *bits*, also known as an 8-bit *byte*.

repertoire.

The collection of characters included in a particular *coded character set*.

scalar value.
A code point expressed as an integer without regard to a particular encoding form; for example, a UTF-8 representation is not appropriate. Scalar values may be displayed in binary, decimal, or hexadecimal notation. Hexadecimal is the most common and is used throughout this document except where binary is required for illustrative purposes.

script.
The set of characters used to write a language. Some scripts serve more than one language.

space (SP).
ASCII code point 20(hex) which is interpreted as a *graphic character* with the unusual property of being recognized in all of the *standard character sets* in the *MARC-8 repertoire* even when not defined in such a set. This *character* is also referred to as "blank" in MARC 21 documentation.

UCS/Unicode.
The Universal Character Set (UCS) embodied in ISO/IEC 10646 and its industry counterpart, Unicode. By design Unicode and ISO/IEC 10646 encode the same character repertoire using identical code points character by character.

UCS/Unicode encoding.
Representation of characters by the code points specified for them in ISO/IEC 10646 and the Unicode Standard. Once established, the code point for a character is unchanging.

UCS/Unicode repertoire.
Over 100,000 characters for all scripts, symbols, and other characters included in ISO/IEE 10646 and the Unicode Standard. Characters continue to be added. The most recent version can be found at www.unicode.org.

Unicode.
see UCS/Unicode

UTF-8.
UCS Transformation Format-8, an encoding form that algorithmically converts Unicode *scalar values* to an *octet*-based format. A particular character in UTF-8 may require from one to four octets. The algorithm is described in Part 3.

working set.
The *coded character set[s]* currently *invoked*.

STANDARDS

- *Character code structure and extension techniques* (ISO/IEC 2022)
- *Code for Information Interchange* (ASCII) (ANSI X3.4)
- *Code Extension Techniques for Use with 7-bit and 8-bit Character Sets* (ANSI X3.41)
- *Coded Arabic Character Set for Information Interchange (ISO 9036) (equivalent to ASMO Standard Specification 449) - except the MARC 21 set contains 5 additional characters and Arabic digits 0-9.*
- *East Asian Character Code for Bibliographic Use* (EACC) (Z39.64)
- *Extended Latin Alphabet Coded Character Set for Bibliographic Use* (ANSEL) (ANSI Z39.47)
- *Extension of the Arabic Alphabet Coded Character Set for Bibliographic Information Interchange* (ISO 11822)
- *Extension of the Cyrillic Alphabet Coded Character Set for Bibliographic Information Interchange* (ISO 5427)
- *Greek Alphabet Coded Character Set for Bibliographic Information Interchange* (ISO 5428)
- *ISO 7-bit coded character set for information interchange* (ISO/IEC 646 (IRV))
- *Hebrew Alphabet Coded Character Set for Bibliographic Information Interchange* (ISO 8957)
- *Universal Multiple-Octet Coded Character Set* (UCS) (ISO/IEC 10646)
- *The Unicode Standard 5.0*, or latest version is found at: www.unicode.org
- *International Register of Coded Character Sets to be Used with Escape Sequences, Registration Number 37, Basic Cyrillic Graphic Character Set*
